



UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

September 21, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES B-11

MEMORANDUM FOR

Howard Hogan
Chief, Decennial Statistical Studies Division

From:

Donna Kostanich *DK*
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by:

MSA Mark Asiala and *MS* Michael Starsinic
Variance Estimation Branch

Subject:

Accuracy and Coverage Evaluation Survey: Variance Estimates by
Size of Geographic Area (Prototype)

The attached document is a prototype of the report that we will prepare, per your request, following completion of applicable Accuracy and Coverage Evaluation Survey (A.C.E.) operations. The completed report is intended to aid the Executive Steering Committee on A.C.E. Policy (ESCAP) in its recommendation regarding the release of the statistically corrected data or the data without statistical correction as the P.L. 94-171 data. This report, together with other reports, will assess the operations and results of both the initial Census and the A.C.E. Both sets of assessments will be available to the ESCAP to aid the Committee in reaching its recommendation regarding the use of the statistically corrected data.

The attached prototype contains both empty table shells and a description of textual analysis that will assess specific aspects of the applicable operations. This report focuses primarily on distributions of coefficients of variation of A.C.E. small-area estimates. In addition to summarizing results for selected subnational geographic areas, it also includes national level results, and, where applicable, compares them to results from the 1990 Post-Enumeration Survey.

It is important to note that the conduct of the operations may lead us to modify the attached format by including additional information. It is also likely that descriptions and definitions will be enhanced or the data items could undergo revision. Conversely, we may conclude, for a variety of reasons, that some of the information set forth in the attached prototype may not be available. The attached document sets forth our conclusions prior to completion of the A.C.E. about what information would properly inform the ESCAP on this subject, but is subject to modification.

Accuracy and Coverage Evaluation 2000: Variance Estimates by Size of Geographic Area

prepared by Mark Asiala and Michael Starsinic

Introduction

The Accuracy and Coverage Evaluation (A.C.E.) consists of two independent samples. The first is a sample of the population in the selected A.C.E. sample areas, known as the P sample. By matching these people to census records, an estimate of the proportion of the population that was missed in the census can be determined. The second is a sample of the census enumerations in the same A.C.E. sample areas, known as the E sample. Using the results of matching the P sample to the census, checking for duplication among the census records and re-interviewing to determine correct inclusion of the census record, an estimate of the proportion of correctly enumerated records in the census can be determined. Dual system estimates (DSE's) are formed for up to 448 post-strata at the national level.

The sampling variance of the A.C.E. is expected to be smaller than the sampling variance of the 1990 Post-Enumeration Survey (PES). In part, this is due to the much larger sample size of the A.C.E. as opposed to the PES: 300,913 housing units in 11,303 clusters for the A.C.E. (Fenstermaker 2000), versus approximately 165,000 housing units in approximately 5,000 clusters for the PES (Obenski & Fay 2000). In general, a sample size increase of this magnitude will decrease the sampling variance. Other factors expected to decrease the sampling error include use of better measures of size in the sample selection and reduction in the variability of sampling weights. Also, instead of a complete search for additional matches and duplicates in blocks surrounding sampled clusters as in the PES, a 20% targeted sample of A.C.E.-selected clusters was chosen to be searched (Targeted Extended Search or TES). The goal of the extended search operation is to reduce the variance of the DSE. The implementation of a more efficient operation relative to the 1990 PES can also result in a reduction of systematic error in the estimates. However, the TES will increase the sampling variance relative to the 1990 PES, since the search will be a sample-based operation. Many of these additional matches and duplicates are highly clustered in a relatively small number of blocks. Because the TES is targeted to clusters with a large number of non-matched or mis-geocoded housing units (which are good indicators of additional matches and duplicates), it will be more efficient than the PES surrounding block search in finding additional matches and duplicates (Navarro 2000).

Overall Assessment

[Prototype Note: Assessment will follow the completion of A.C.E. variance estimation.]

Variance Estimation Methodology

The A.C.E. survey was a multi-phase sample, which increased the difficulties of estimating the

sampling variance. Multi-phase sampling differs from multi-stage in the following way: in a multi-stage design, the information needed to draw all stages of the sample is known before the sampling begins; in a multi-phase design, the information needed to draw the n^{th} phase of the sample is unobtainable until the $n-1^{\text{st}}$ phase of the sample is completed. A methodology based in part on the Rao-Shao jackknife variance estimator (Rao & Shao 1992) takes into account the multi-phase nature of the A.C.E. The estimation of the variance due to the A.C.E. attempts to capture these components of the variance:

- Sampling variance due to the initial Listing sample.
- Sampling variance due to the A.C.E. Reduction and Small Block Subsampling.
- Sampling variance due to the Targeted Extended Search (TES) sample.
- Variance due to the imputation of correct enumeration, match, and residence probabilities for unresolved cases.

Variances are directly estimated by this new methodology only for the final collapsed post-strata (up to 448 post-strata), and a variance-covariance matrix for the Coverage Correction Factors (CCF) is created. (The relative contribution to the sampling error from the above four components is not considered in this analysis.) The estimated (“synthetic”) variance of any population estimate can be computed using this matrix and the uncorrected census counts, broken down by post-stratum and excluding persons out-of-scope of the A.C.E. (For more information see Kim et al, 2000, and Starsinic & Kim, 2000.)

$$\begin{aligned}\hat{X}_s &= \text{Synthetic household population estimate for geographic area } s \\ &= \sum_{\text{post-strata } h} \hat{X}_{sh} \\ &= \sum_{h=1}^H C_{sh} \times \text{CCF}_h, \text{ where } C_{sh} = \text{Census count of post-stratum } h \text{ in geographic area } s\end{aligned}$$

$$\begin{aligned}\text{Var}(\hat{X}_s) &= \text{synthetic variance for synthetic household population estimate } \hat{X}_s \\ &= \text{Var}\left(\sum_{h=1}^H \hat{X}_{sh}\right) \\ &= \sum_{h=1}^H \sum_{h'=1}^H \text{Cov}(\hat{X}_{sh}, \hat{X}_{sh'}) \\ &= \sum_{h=1}^H \sum_{h'=1}^H \text{Cov}(C_{sh} \times \text{CCF}_h, C_{sh'} \times \text{CCF}_{h'}) \\ &= \sum_{h=1}^H \sum_{h'=1}^H C_{sh} \times C_{sh'} \times \text{Cov}(\text{CCF}_h, \text{CCF}_{h'})\end{aligned}$$

This estimate of variance is only intended to include the error from the above four components, and is not intended to quantify nonsampling errors, other than A.C.E. imputation error. One component of the total error which is specifically not incorporated into this calculation is the “synthetic” (or model) error. This model assumes that the coverage rate is uniform over all areas within post-strata. To the extent that areas deviate from this assumption we are introducing synthetic or model error. The accuracy of this methodology may decrease in areas where localized effects not reflected in the post-stratification affect the true sampling variance. This discrepancy becomes more pronounced as the population of an area decreases. Thus, caution should be used in comparisons between areas of different sizes. The intent is to see how A.C.E. and PES sampling variances compare.

Components of sampling error which are not incorporated into the variance estimates are the error due to weight trimming, and the error due to large block subsampling.

Measuring Reliability

Variances are useful associated with an estimate, but lose their utility when taken out of context. A variance of 5 implies different things when the estimate is 500 as opposed to an estimate of 0.5. A useful measure of an estimate’s “reliability” is the coefficient of variation (CV), defined as the ratio of the square root of the variance of an estimate and expected value of the estimate. The smaller the value for the CV, the more “precise” the estimator can be considered to be. We will concentrate on the CV of the synthetic (small-area) estimate. For any desired population estimate, geographic or otherwise:

$$\text{Synthetic total population estimate} = \text{Synthetic household population estimate } (\hat{X}) + \text{“Residual” count}$$

where the “Residual” count are persons out-of-scope of the A.C.E. sample. These include institutionalized and non-institutionalized group quarters persons; persons counted in Service Based Enumeration (SBE), and those estimated by the SBE’s multiplicity estimator; and persons enumerated in the Remote Alaska operation. (Variance due to the SBE’s multiplicity estimation is not accounted for in the A.C.E. variance estimates.) Note that this analysis is limited to preliminary results of small-area estimates and variance estimates. These preliminary small-area estimates used in this report will not have undergone controlled rounding and therefore will not agree with forthcoming official published results. However, the effect of controlled-rounding on the final population estimates is negligible and for the purpose of this analysis the differences will have no effect on the conclusions. The CV is computed as:

$$CV = \frac{\sqrt{\text{Var}(\text{Synthetic total population estimate})}}{\text{Synthetic total population estimate}}$$

Since the Residual population is excluded from the A.C.E. sample, it adds no sampling variance, and the variance of the synthetic estimate is the same as the variance of the corresponding A.C.E.

estimate described above.

In addition to the reliability of the population estimates themselves, the reliability of the net undercount percent (UC) is also of interest.

$$UC = 100\% \times \frac{\text{Synthetic total population estimate} - \text{Census count}}{\text{Synthetic total population estimate}}$$

The standard error of the net undercount percent can easily be approximated based on the CV of the synthetic estimate:

$$SE(UC) = (100\% - UC) \times CV$$

Results

Table 1 provides national summary statistics from both the 2000 synthetic estimates and the 1990 PES estimates, for states, congressional districts, places with a Census population greater than 100,000 (determined separately on 1990 and 2000 data), and counties with a Census population greater than 100,000 (determined separately on 1990 and 2000 data). The 1990 PES data are based on the final 357 post-strata definitions, and come from Thompson, 1992. Standard descriptive statistics for the CVs are given including N (number of geographic areas within the United States), mean, minimum, maximum, median, first and third quartiles. In addition, the mean size (in population) of the geographical area is given, and ME denotes the 90% margin of error for a geographical area of size equal to the mean size with a CV equal to the mean CV.

$$ME = 1.645 \times \text{Synthetic Population Estimate} \times CV$$

Table 2 provides direct comparisons between 1990 PES and 2000 synthetic estimates for states. The 1990 data provided are 1990 census count (Census), 1990 PES population estimate using the final 357 post-strata definitions (357 PES Est), CV of the synthetic estimate (CV), net undercount percent (UC), and standard error of the net undercount percent (SE(UC)). The 2000 data provided are 2000 census count (Census), 2000 synthetic estimate (Syn Est), CV of the synthetic estimate (CV), net undercount percent (UC), standard error of the net undercount percent (SE(UC)), and the margin of error (ME) as defined above.

Graphs 1a & b, 2a & b, and 3a & b depict the distribution of CV's of the 1990 ("a" graphs) and 2000 ("b" graphs) data for congressional districts, places with a Census population greater than 100,000, and counties with a Census population greater than 100,000, with these last two categories defined as in Table 1. [Prototype Note: Only "a" graphs are depicted, as 2000 data is not yet available. The "b" graphs' layouts will be identical to their "a" counterparts.]

Graph 4a and 4b show the 1990 and 2000 net undercount percent estimates with 90% confidence intervals, by state, respectively. [Prototype Note: As this graph is for demonstration purposes

only, it does not include further pages with the remaining 34 states. The production version obviously will.]

[Prototype Note: Since the vast majority of analysis which will be included in this report concerns comparisons between the PES and the A.C.E., more detailed analyses are not possible without the actual A.C.E. data.]

References

Fenstermaker, D., DSSD Census 2000 Procedures and Operations Memorandum Series R-33, "Accuracy and Coverage Evaluation Survey: Sample Design Summary".

Haines, D., DSSD Census 2000 Procedures and Operations Memorandum Series Q-30, "Accuracy and Coverage Evaluation Survey: Computer Specifications for Person Synthetic Estimation (U.S.)".

Kim, J.K., Navarro, A., and Fuller, W., "Replication Variance Estimation for Multi-Phase Stratified Sampling", Unpublished Census Bureau Memorandum, July 24, 2000.

Navarro, A., DSSD Census 2000 Procedures and Operations Memorandum Series Q-18, "Accuracy and Coverage Evaluation Survey: Targeted Extended Search Plans".

Obenski, S., and Fay, R., "Analysis of C.A.P.E. Findings on PES Accuracy at Various Geographic Levels", internal Census memorandum, June 6, 2000.

Rao, J.N.K., and Shao, J., "Jackknife variance estimation with survey data under hot deck imputation." *Biometrika*, 79, 811-812, 1992.

Starsinic, M., and Kim, J.K., DSSD Census 2000 Procedures and Operations Memorandum Series V-2, "Computer Specifications for Variance Estimation for Census 2000".

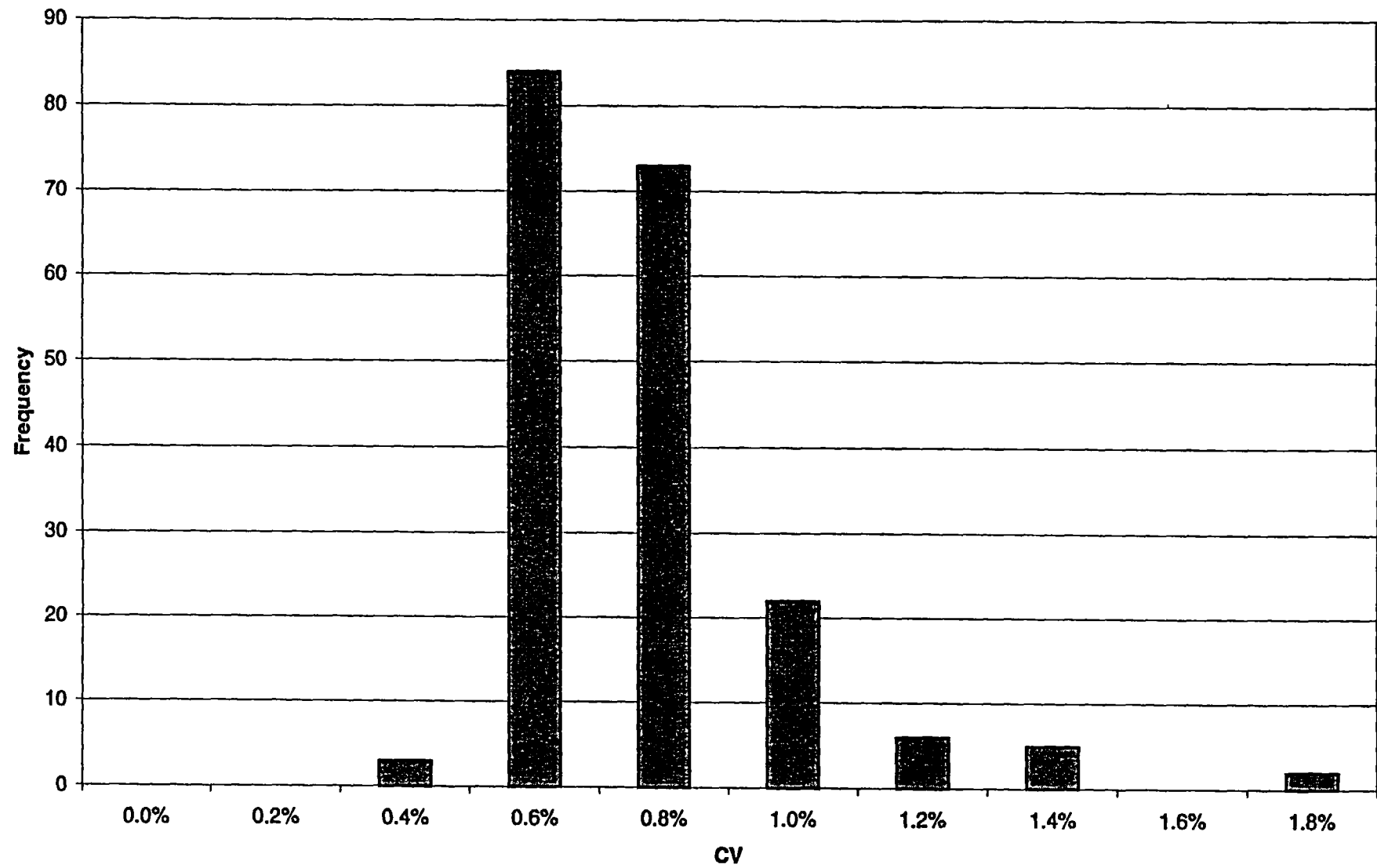
Thompson, J., "357 Post Enumeration Survey (PES) Estimates". Internal census memorandum to the CAPE Committee and CAPE Working Group, April 24, 1992.

**Table 1: US Summary of Distribution of CVs by Geographical Area
for 1990 PES and 2000 A.C.E**

[illegible]

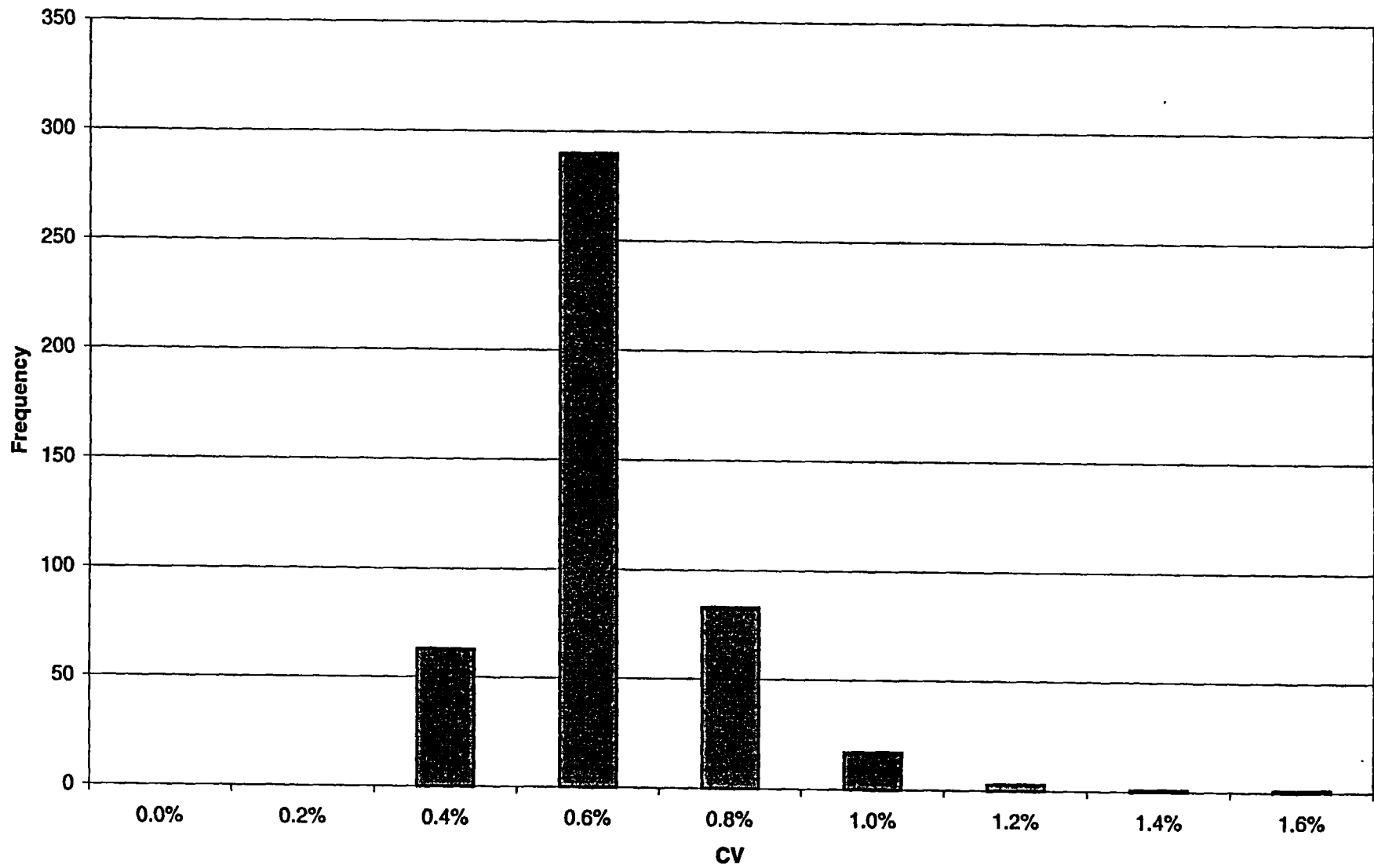
Graph 2a:

Distribution of CV's for Places with Population Greater Than 100,000 for 1990 PES



Graph 3a:

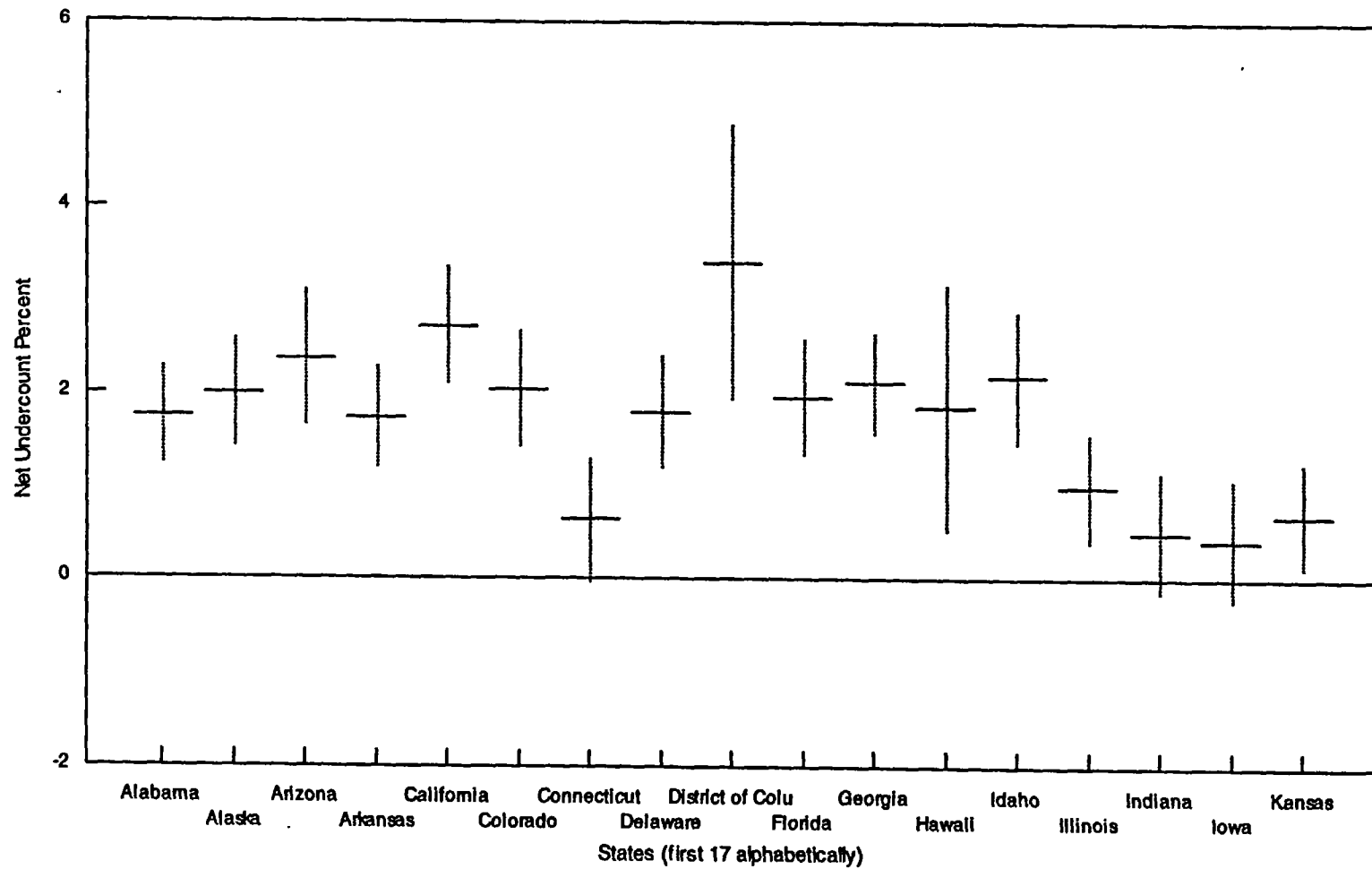
Distribution of CV's for Counties with Population Greater Than 100,000 for 1990 PES



Graph 4a:

1990 PES Net Undercount Percent Estimates, with 90% Confidence Intervals

First 17 States (alphabetically)



— 90% Confidence Interval

— Net Undercount Percent Estimate